

Visualizing Associations Between Genome Sequences and Gene Expression Data Using Genome-Mean Expression Profiles

Derek Y. Chiang

Patrick O. Brown

Michael B. Eisen

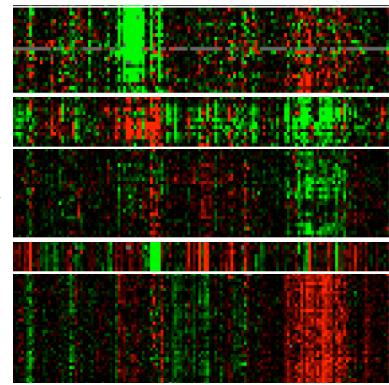
ISMB 2001

ISMB 2001 – Chiang D.Y., et al.

Regulation of Gene Expression

What information controlling gene expression is encoded in genome sequences?

```
>Saccharomyces cerevisiae chr V
CGTCTCTCCAAGCCCTGGTGTCTTACCC
GGATGTTCAACAAAAGCTACTTACTACCTT
TATTTTATGTTACTTTTATAGATTGTCTT
TTTATCCTACTCTTCCCCTGTCTCTCGC
TACTGCCGTGCAACAAACACTAAATCAAAC
AGTGAAATACTACTACATCAAACGCATATT
CCCTAGAAAAAAAATTCTTACAATATACT
ATACTACACAATACATAATCACTGACTTTCG
TAACAACAATTCTCTTCACTCTCAACTCT
CTGCTGAATCTCTACATAGTAATATTATAT
CAAATCTACCGTCTGGAACATCATCGCTATC
CAGCTTTGTGAACCGCTACCATCAGCATG
TACAGTGGTACCTTCGTGTTATCTGCAGCGA
GAACCTCAACGTTGCCAAATCAAGCCAATG
TGGTAACAAACCACACCTCCGAAATCTGCTCC
AAAAGATACTCCAGTTCTGCCGAAATGTTT
```



ISMB 2001 – Chiang D.Y., et al.

Regulation of Gene Expression

PROBLEM

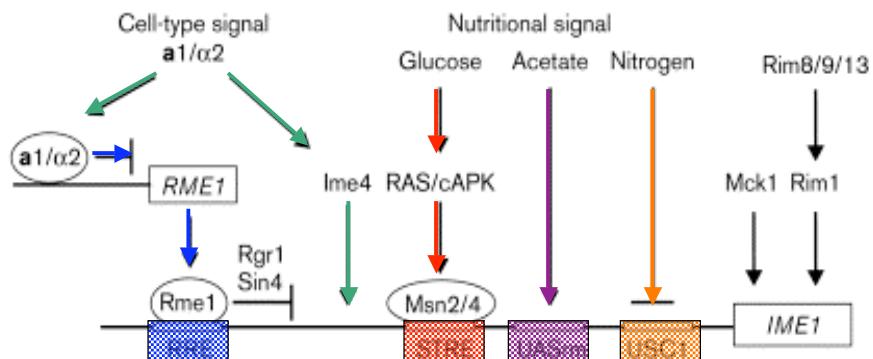
- Find **DNA sequences** in promoters that are recognized by **transcription factors**



ISMB 2001 – Chiang D.Y., et al.

Regulation of Gene Expression

Combinatorial Regulation

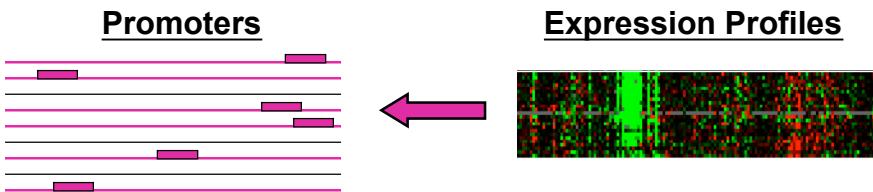


REF: Vershon, A. K. & Pierce, M. (2000)
Curr. Opin. Cell Biol. 12:334-339

ISMB 2001 – Chiang D.Y., et al.

Current Computational Methods

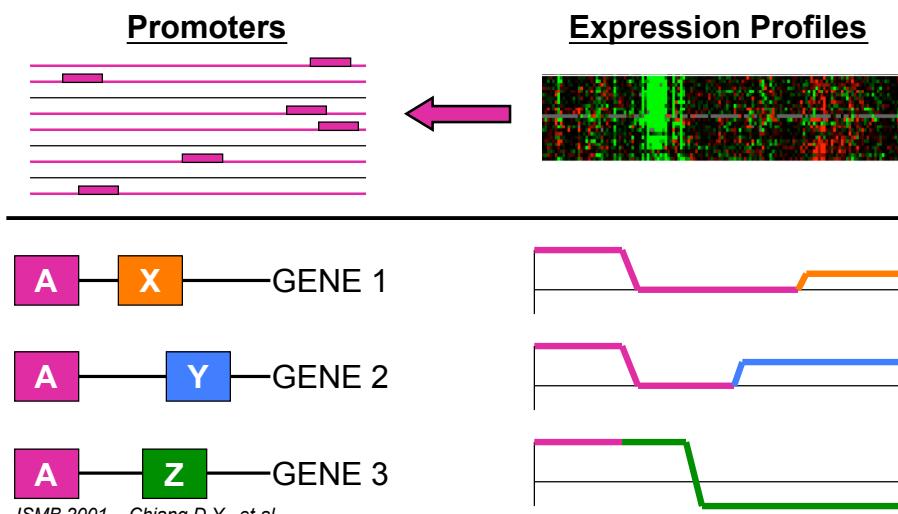
Group-by-EXPRESSION Approach



ISMB 2001 – Chiang D.Y., et al.

Current Computational Methods

Group-by-EXPRESSION Approach



ISMB 2001 – Chiang D.Y., et al.

An Alternative Approach

Group-by-SEQUENCE Approach

- Do expression profiles of genes with shared promoter sequences reflect TF activity?

Promoters



Expression Profiles



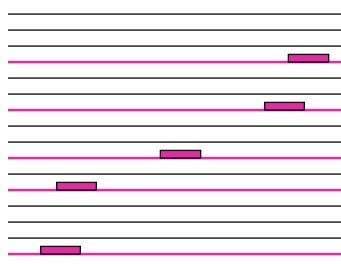
ISMB 2001 – Chiang D.Y., et al.

An Alternative Approach

Group-by-SEQUENCE Approach

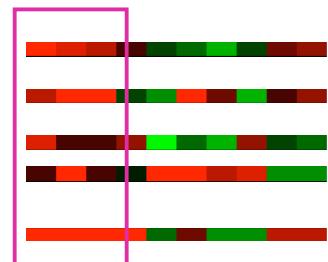
- Do expression profiles of genes with shared promoter sequences reflect TF activity?

Promoters



Expression Profiles

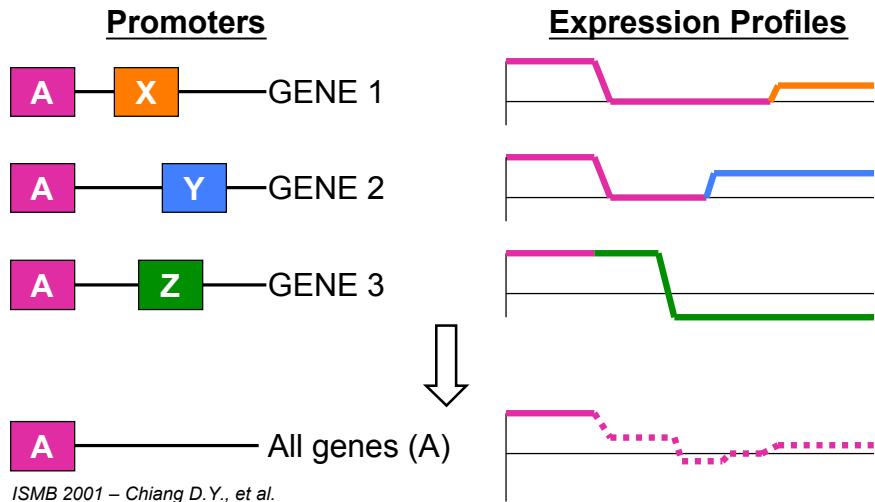
Binding site



ISMB 2001 – Chiang D.Y., et al.

Genome-Mean Expression Profiles

Quantifying “Shared Features”

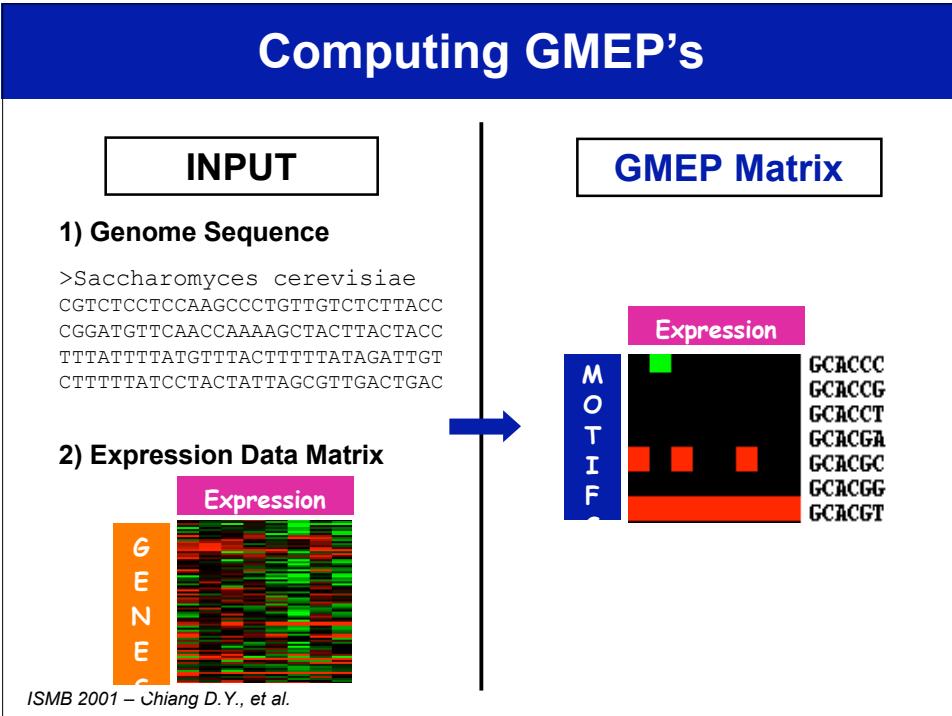


Genome-Mean Expression Profiles

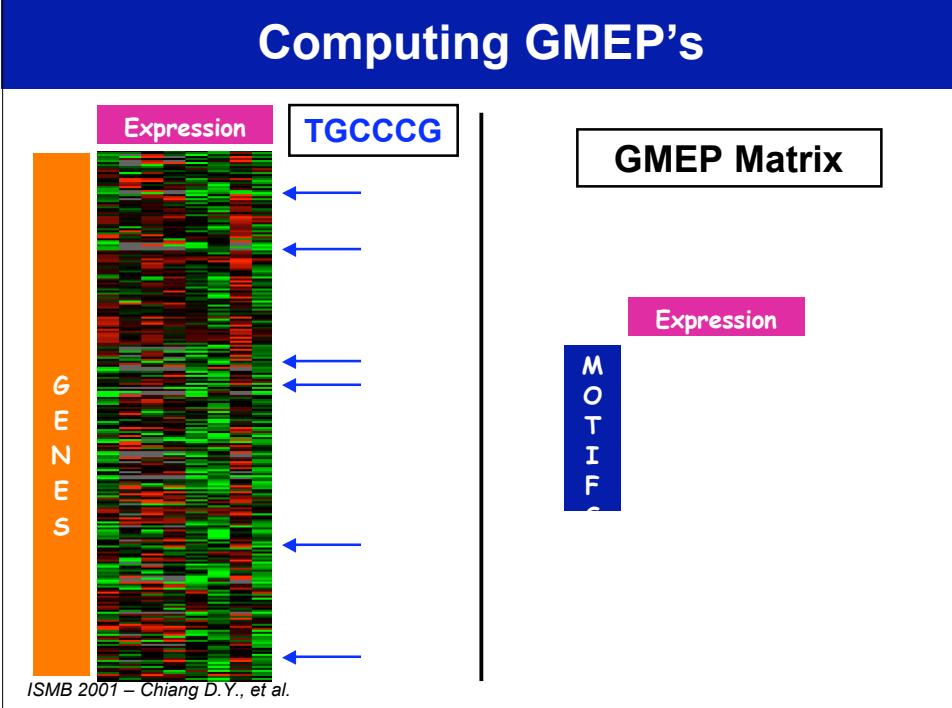
- Computing Genome-Mean Expression Profiles
- Statistical Significance
- Biological Relevance
- Find Transcription Factor Binding Sites

ISMB 2001 – Chiang D.Y., et al.

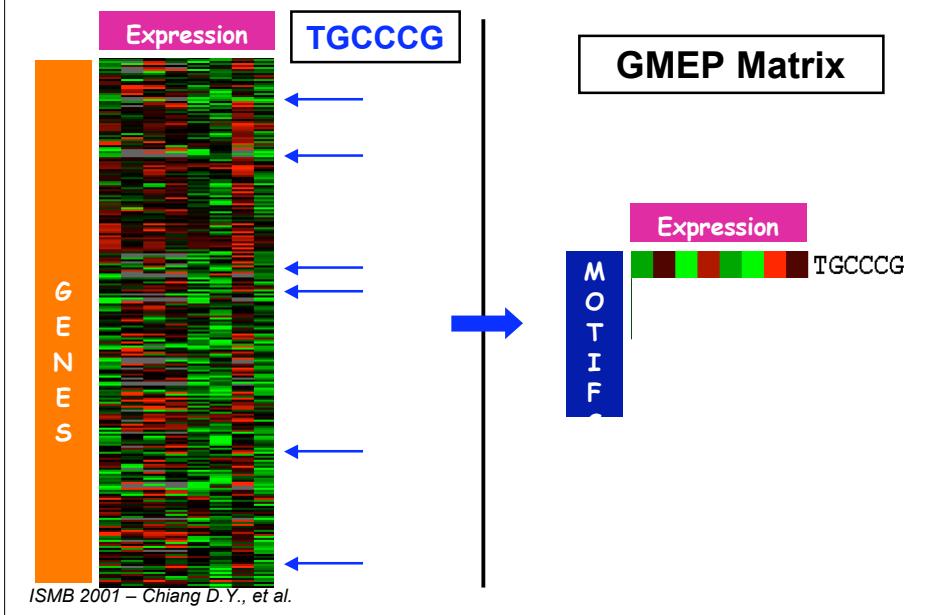
Computing GMEP's



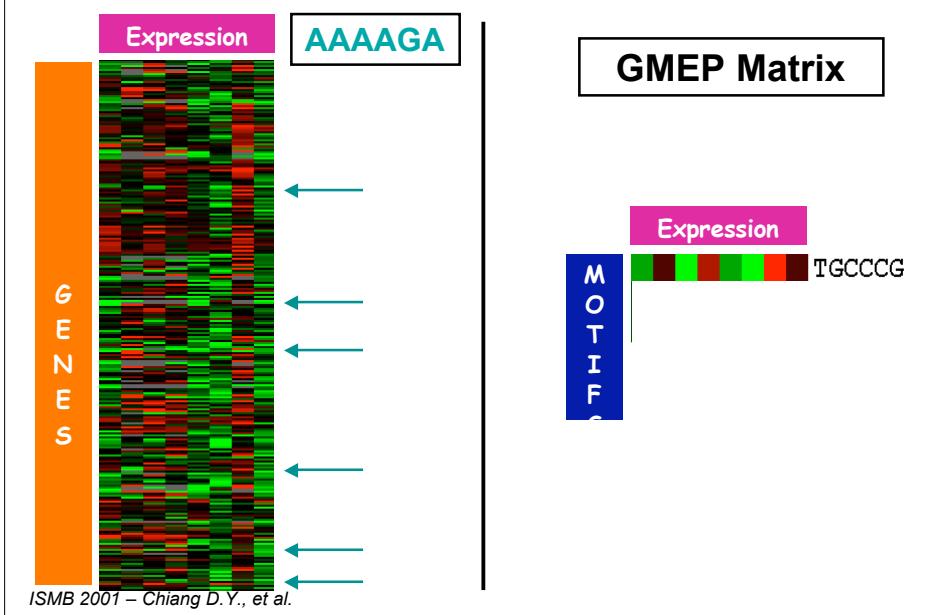
Computing GMEP's



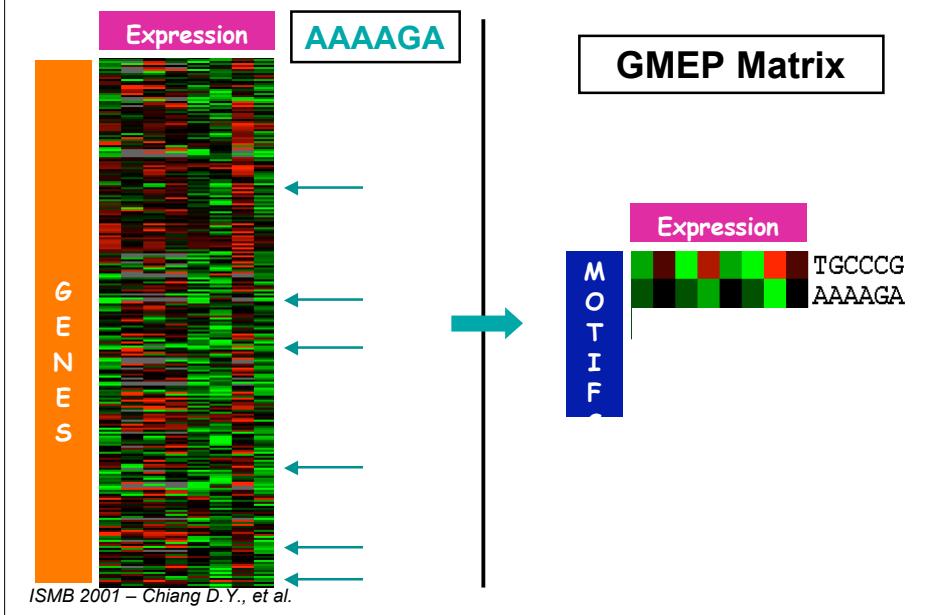
Computing GMEP's



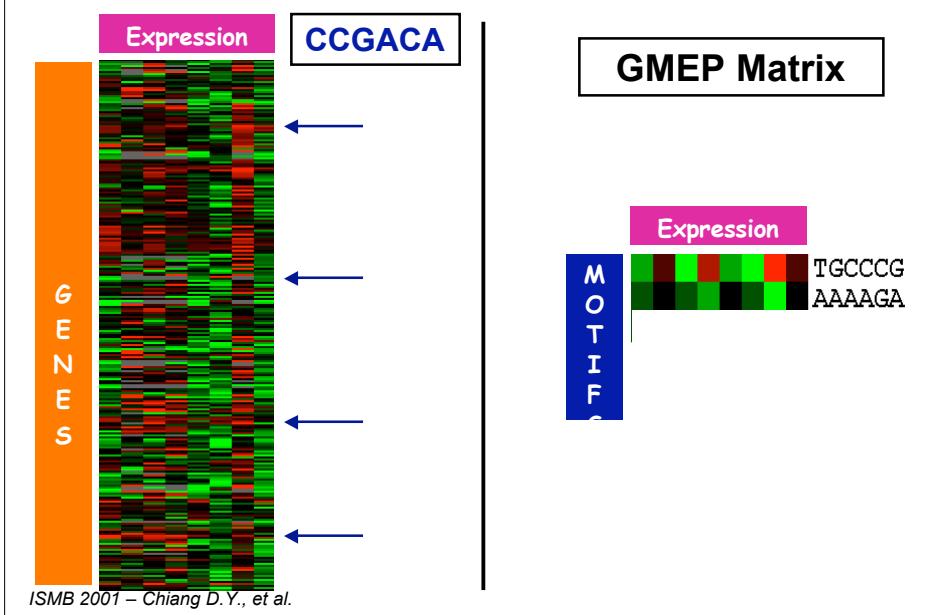
Computing GMEP's



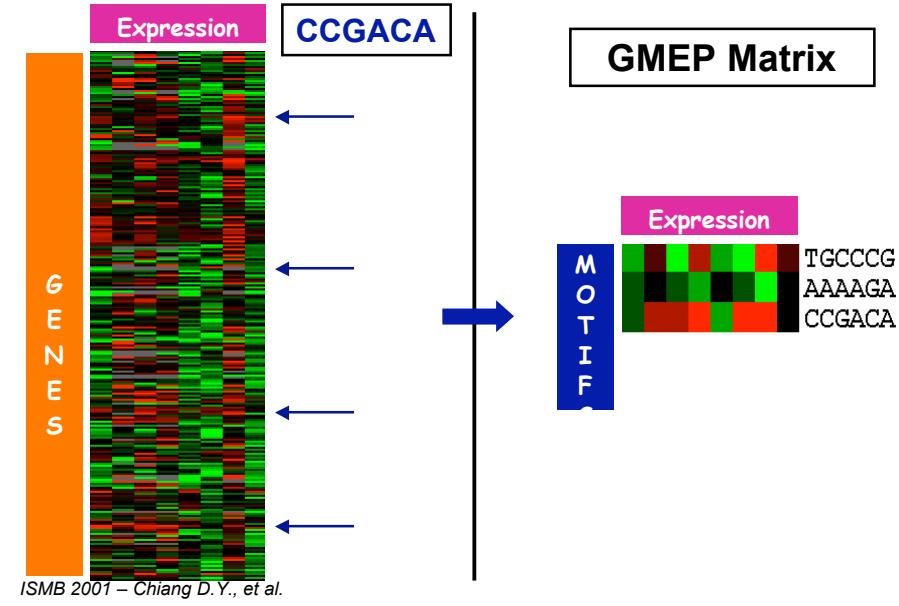
Computing GMEP's



Computing GMEP's



Computing GMEP's



Computing GMEP's

gmep implemented in C++ on RedHat Linux 6.1
Download from <http://rana.lbl.gov>

ALGORITHM

Read input gene expression data

```

1: for gene 1 to 5300
2:   for pos 1 to ( $U - L$ )
3:      $m = \text{Genome}(\text{pos}) \dots \text{Genome}(\text{pos} + L)$ 
4:     for experiment e 1 to  $E$ 
5:        $S[m][e] += D[\text{gene}][e]; W[m][e]++$ 
6:     end for
7:   end for
8: end for

```

Compute GMEP's

```

9: for  $m$  1 to  $4^L$ 
10:   for experiment e 1 to  $E$ 
11:      $GMEP[m][e] = S[m][e] / W[m][e]$ 
12:   end for
13: end for

```

Runtime: $O((4^L + gU) \times e)$

ISMB 2001 – Chiang D.Y., et al.

Genome-Mean Expression Profiles

GMEP Approach

- Computing GMEP's
- **Statistical Significance**
- Biological Relevance
- Find Transcription Factor Binding Sites

ISMB 2001 – Chiang D.Y., et al.

Statistical Significance: Z-scores

Z score

$$Z(GMEP(m)_j) = \frac{GMEP(m)_j - E(GMEP)_j}{SD(GMEP(m)_j)}$$

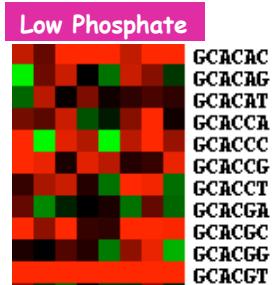
-
- Correct for varying # of occurrences of motifs:

$$Z(GMEP(m)_j) = \frac{GMEP(m_j) - \bar{I}_j}{\sigma_j / \sqrt{N_m}}$$

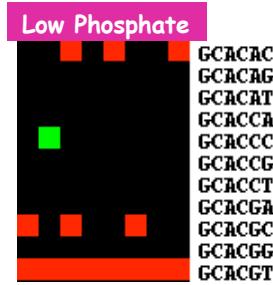
ISMB 2001 – Chiang D.Y., et al.

Statistical Significance: Z-scores

No Filter



Filter: $Z > 3$



ISMB 2001 – Chiang D.Y., et al.

Genome-Mean Expression Profiles

GMEP Approach

- Computing GMEP's
- Statistical Significance
- **Biological Relevance**
- Find Transcription Factor Binding Sites

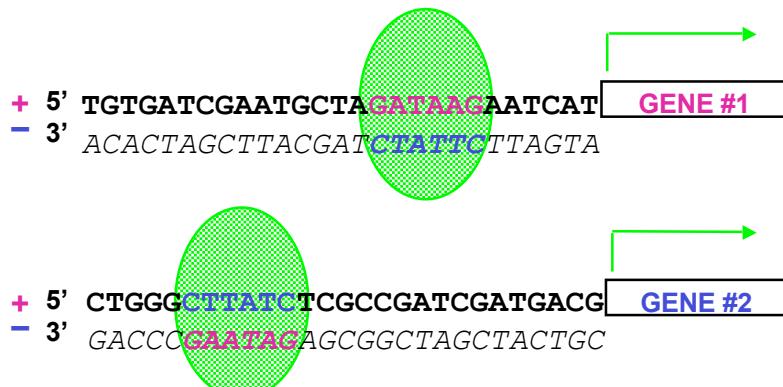
ISMB 2001 – Chiang D.Y., et al.

Biological Relevance

Transcription Factor – DNA Interactions

- Recognize both DNA strands
- Orientation-independent

5' **GATAAG** 3'
3' **CTATTC** 5'



ISMB 2001 – Chiang D.Y., et al.

Cross-Validation of Binding Sites

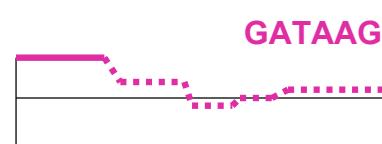
- Only + strand was used to compute GMEP's

Promoters

+ GGCTCC**GATAAG**CTTTTGCG---**GLT1**
- CCGAGGCTATTGAAACG

+ **AGATAAGATAACAAATCAT**---**GAT1**
- TCTATTCTATTGTTAGTA

GMEP



ISMB 2001 – Chiang D.Y., et al.

Cross-Validation of Binding Sites

TF Binding Site:



Promoters

+ GGCTCC**GATAAG**CTTTTGC---**GLT1**

- CCGAGGCTATTGAAAAACG

+ **A**GATAAGATAACAAATCAT---**GAT1**

- TCTATTCTATTGTTTAGTA

+ GATCCTGGG**CTTATC**TCGC---**DAL80**

- CTAGGACCCGAATAGAGCG

+ **ATTTCC**CTTATC**ATCTCAT**---**PUT1**

- TAAAGGGAATAGTAGTGTA

GMEP

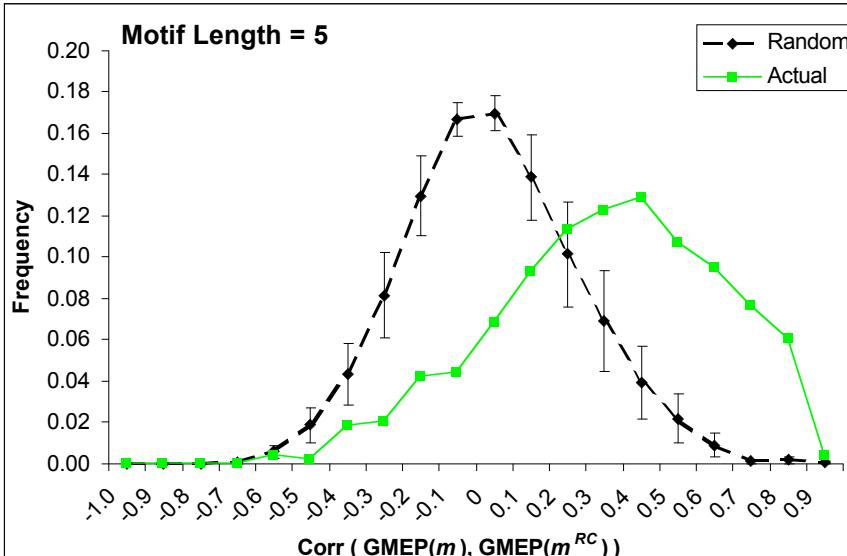
GATAAG



CTTATC

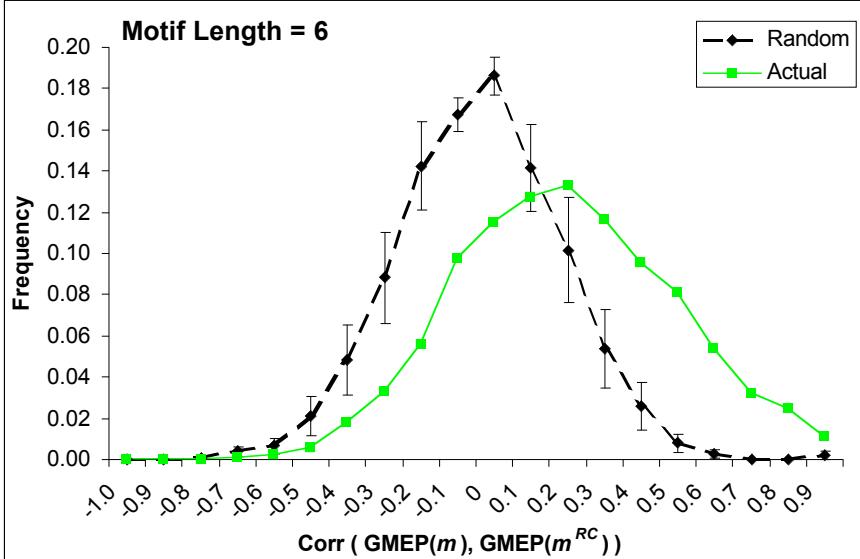
ISMB 2001 – Chiang D.Y., et al.

GMEP Correlations with Reverse Complement



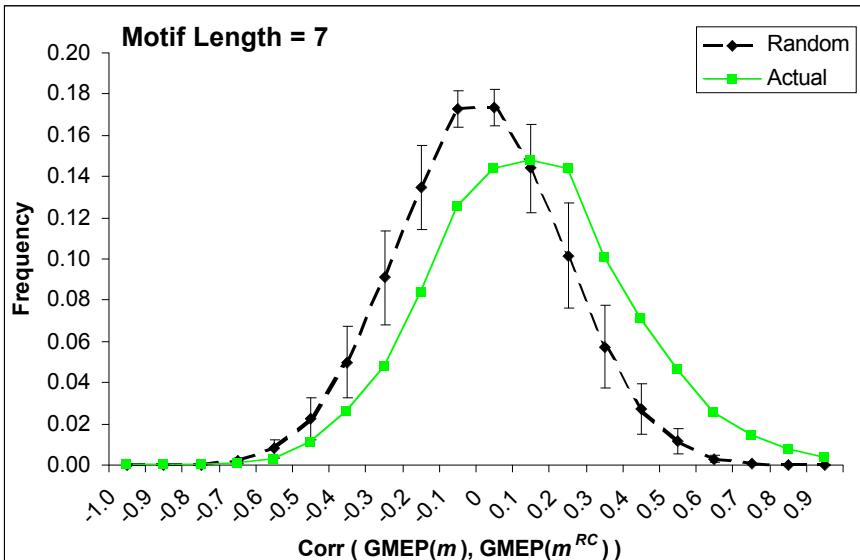
ISMB 2001 – Chiang D.Y., et al.

GMEP Correlations with Reverse Complement



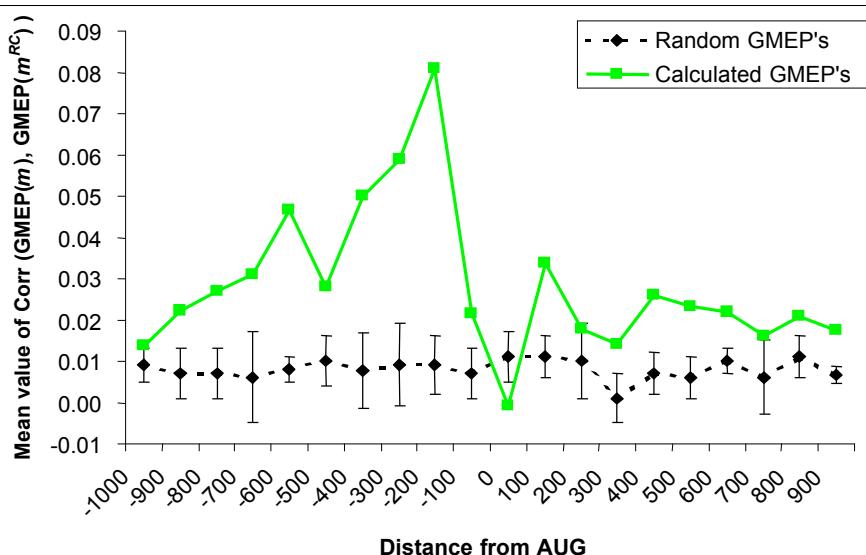
ISMB 2001 – Chiang D.Y., et al.

GMEP Correlations with Reverse Complement



ISMB 2001 – Chiang D.Y., et al.

Position Effects of GMEP Correlations



ISMB 2001 – Chiang D.Y., et al.

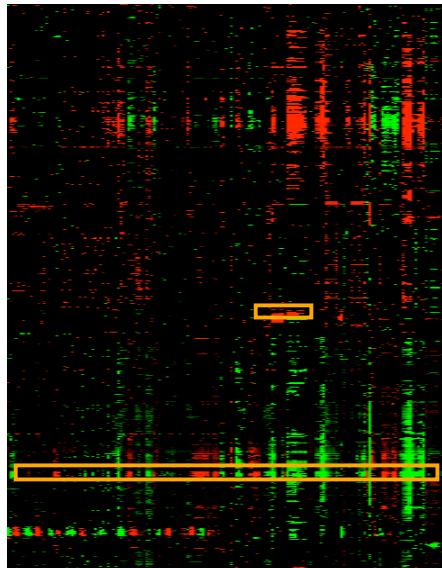
Genome-Mean Expression Profiles

GMEP Approach

- Computing GMEP's
- Statistical Significance
- Biological Relevance
- **Find Transcription Factor Binding Sites**

ISMB 2001 – Chiang D.Y., et al.

Identify TF Binding Sites



GMEP Matrix for SMD

Filter: $Z > 3$

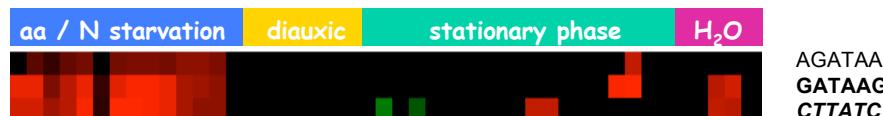
► Example #1

► Example #2

ISMB 2001 – Chiang D.Y., et al.

Identify TF Binding Sites

Condition-Specific Binding Site



NIT: Nitrogen-regulated genes

5' **GATAAG** 3'

REF: van Helden, J., et al. (1998) *J. Mol. Biol.* **281**: 827-842

Nitrogen metabolism DAL1, DAL2, DAL3, DAL4, DAL5, DAL7

Nitrogen uptake DUR3, MEP1, MEP2, GAP1

Transcription factors DAL80, DAL81, DAL82, GAT1, GLN3, GZF3, NPR2

ISMB 2001 – Chiang D.Y., et al.

Identify TF Binding Sites

Multiple-Condition Binding Site



Stress Repressive Element 5' TG^C_A GATGAG^C_A T 3'

REF: Gasch, A. P. et al. (2000) *Mol. Biol. Cell* 11: 4241-4257

rRNA processing

DBP10, DRS1, MTR3, MTR4, RRP6, SPB4, etc.

Ribosome

RPL1A, RPL7A, RPL25, RPS3, etc.

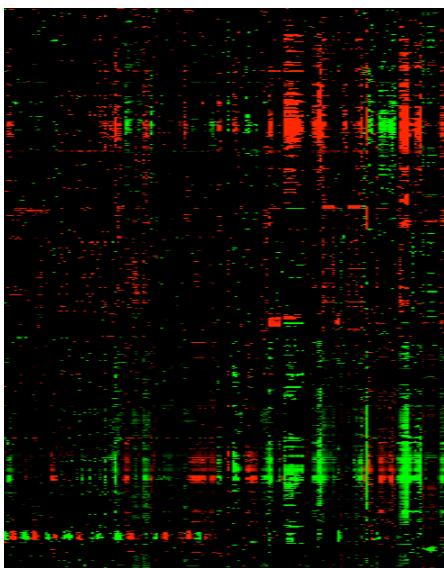
Unknown function

KRE33, NAN1, NOP14, PWP1, etc.

ISMB 2001 – Chiang D.Y., et al.

Identify TF Binding Sites

Known Binding Sites

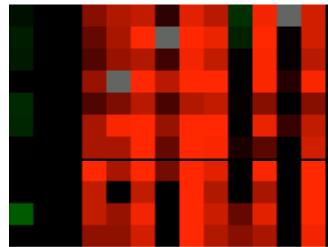


■ Msn2/4p	CCCCT
■ Pho4p	CACGTG
■ Rpn4p	^C GTGGCAAA _G
■ Gln3p	GATAAG
■ Gcn4p	TGA ^C _G TCA
■ Hsf1p	TTCTCGAA
■ Unknown	TGAAAATTTT
■ Unknown	TG ^C _A GATGAG ^C _A T
■ Mbf1p	^A CGCG ^A _T

ISMB 2001 – Chiang D.Y., et al.

Identify TF Binding Sites

Ergosterol



ERG6
ERG28
ERG26
ERG3
ERG27
ERG5
ERG1
CGTTTA
AACGAG
TCGTTT
ATCAGC

}

ERG Genes

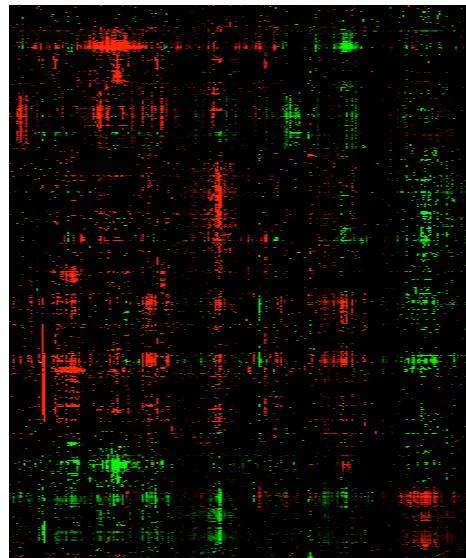
Upc2p site

ERG: Ergosterol-regulated genes 5' CTCGT^A T^A C^A AGC 3'

REF: Vik, A. and Rine, J. (2001) *Mol. Biol. Cell* (in press)

ISMB 2001 – Chiang D.Y., et al.

Identify TF Binding Sites



Gcn4p	TGA ^C TCA _G
Upc2p	CTCGT ^A T ^A C ^A AGC
Ste12p	TGAAACA
Pho4p	CACGTG
Aft1p	TGCACCCGA
Msn2/4p	CCCCT
Rap1p	GCACCC
Pdr1/3p	TCCGCGGA
Mig1p	ACCCCGC
Mbf1p	^A CGCG ^A _T
Unknown	TG _A GATGAG ^C T
Unknown	TTTGAAA

ISMB 2001 – Chiang D.Y., et al.

Acknowledgements

UC Berkeley

Audrey Gasch

Ben Berman

Terry Speed

Mark van der Laan

Åshild Vik

Stanford

Stanford Microarray Database

<http://genome-www.stanford.edu/microarray>

U.S. Department of Energy

National Sciences and Engineering Research Council (Canada)

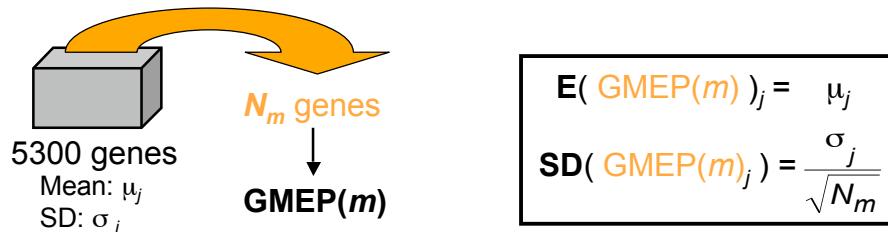
Howard Hughes Medical Institute

ISMB 2001 – Chiang D.Y., et al.

Statistical Significance: Z-scores

Central Limit Theorem approximates $\text{SD}(\text{GMEP}(m))$

- Consider $\text{GMEP}(m)$ as the sample mean of N_m iid gene expression measurements



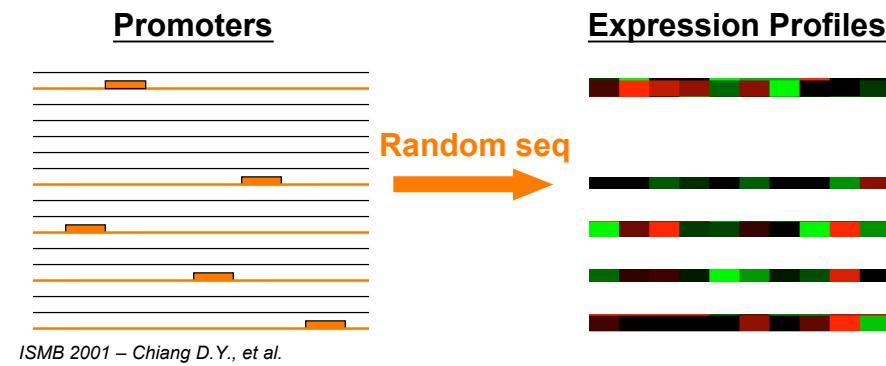
$$Z(m)_j = \frac{\text{GMEP}(m)_j - \mathbf{E}(\text{GMEP})_j}{\mathbf{SD}(\text{GMEP}(m)_j)} = \frac{\text{GMEP}(m)_j - \bar{I}_j}{\bar{\sigma}_j / \sqrt{N_m}}$$

ISMB 2001 – Chiang D.Y., et al.

An Alternative Approach

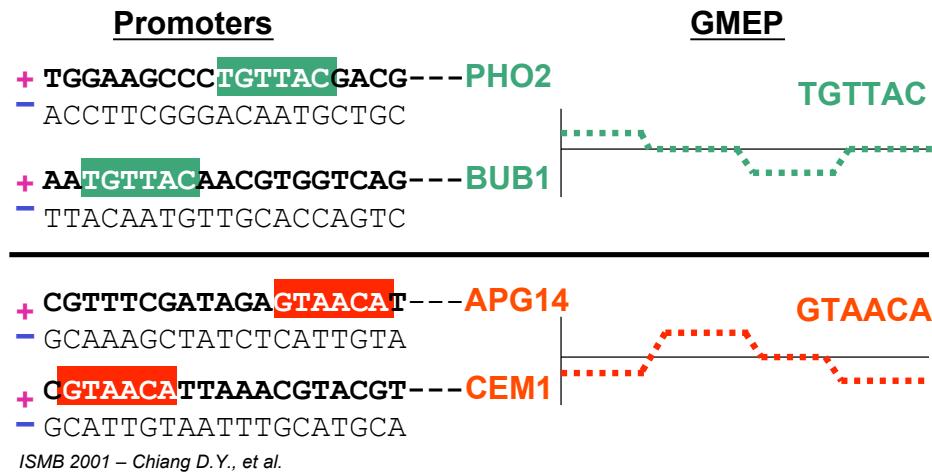
Group-by-SEQUENCE Approach

- Do expression profiles of genes with shared promoter sequences reflect TF activity?



Cross-Validation of Binding Sites

- Only + strand was used to compute GMEP's



Identify TF Binding Sites



PDR: Pleiotropic Drug Resistance

5' **TCCGCGGA** 3'

REF: Hallstrom, T. C. & Moye-Rowley, W. S. (2000) *J. Biol. Chem.* **275**: 37347-37356

ISMB 2001 – Chiang D.Y., et al.